*CS555 Assignment 1*

*Zuowen Tang*

*Feb 2, 2023*

1. Save the data to a CSV file and read into R for analysis. (2 points) ✅

(2) Make a histogram of the duration of days of hospital stays. Ensure the histogram is labelled appropriately. Use a width of 1 day. Describe the shape, center, and spread of the data. Are there any outliers? (5 points)

Chart, histogram

Description automatically generated

Shape: skewed to the right, which means most of the cases fall on the lower end.

The mean and median are very close (5.63 and 5.) The Spread of the data is from 2 to 15.

There is no outlier.

(3) Find the mean, median, standard deviation, first and third quartiles, minimum and maximum of the durations of hospital stay in the sample. Summarize these values in a table that you create in EXCEL or WORD. In other words, do \*not\* simply copy and paste R output. You should be reporting a nicely labeled and formatted table.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mean | Median | Stand. Dev. | 1st Quartiles | 3rd Quartiles | Minimum | Maximum |
| 5.63 | **5** | **2.74379** | **4** | **7** | **2** | **15** |

- Given the shape of the distribution, what is the best single number summary of the center of the distribution?

The Median would be the best single-number summary because most of the cases fall on the right side of the diagram. So the median is closer to the center part.

- What is the best single number summary of the spread of the distribution? (5 points)

The Interquartile Range would be the best single-number summary here because how widely the data spread by using IQR, we can tell the percentile of each interquartile of the data.

(4) Assume that the literature on this topic suggests that the distribution of days of hospital stay are normally distributed with a mean of 5 and a standard deviation of 3. Use R to determine the probabilities below based on the normal distribution described above (you should not be using the data set given on the following page):

1. What percentage of patients are in the hospital for less than 10 days? (4 points)

Chart

Description automatically generated

Based on the above diagram, we can see that the percentage of patients who stayed in the hospital for less than ten days is around 95%.

(b) Recent publications have indicated that hypervirulent strains of C. Difficile are on the rise. Such strains are associated with poor outcomes, including extended hospital stays. An

investigator is interested in showing that the average hospital stay duration have increased

versus published literature. He has a sample of 35 patients from his hospital. If the published

data are consistent with the truth, what is the probability that the sample mean in his sample

will be greater than 6 days? (4 points)

Chart, histogram

Description automatically generated

Above is a histogram to visualize the sampling distribution (a total of 10000 samples containing 35 patients with given mean and standard deviation.) The mean of the sampling distribution is 4.994344, which is very close to the given mean of 5. Under multiple runs, the probability of that sample mean being greater than or equal to 6 is roughly around 0.024 (2.4%.)